SPENDING A RAINY AFTERNOON WITH SAS - NO. 1

Fitting Weighted Linear Regressions with the Statistical Analysis System (SAS)

W. D. Bell
Cornell University

BU-527-M August, 1974

Abstract

These few words are meant to point out how the Statistical Analysis System (SAS) may be employed to fit univariate linear regression models where the error variance is a (known) function of the independent variables. Note that the same principles may be applied in fitting non-linear regressions where the original model may be "transformed" to a linear one and the error variance is a function of the independent variables. This mimeograph ends with a simple programme segment which implements this technique in SAS.

SPENDING A RAINY AFTERNOON WITH SAS - NO. 1

Fitting Weighted Linear Regressions with
the Statistical Analysis System (SAS)

W. D. Bell
Cornell University

BU-527-M August, 1974

Let $y^*: R^k \to R$ be defined by

$$y^*(\underset{\sim}{x}) = f(\underset{\sim}{x};\ \beta_1,\ldots,\beta_m),\ \underset{\sim}{x} \in R^k\ ,$$

with f a linear function of $\beta_1,\ldots,\beta_m$ for every fixed $\underset{\sim}{x}$, and $\{\beta_1,\ldots,\beta_m\}$ a linearly independent set of (unknown) parameters. Let $Y(\underset{\sim}{x}) = y^*(\underset{\sim}{x}) + \epsilon$ where ϵ is a (real) random variable with mean 0 and finite variance.

Suppose one observes the sample $(Y_1,\ldots,Y_n)$ corresponding to $(\underset{\sim}{x}_1,\ldots,\underset{\sim}{x}_n)$. Let $d_i(\beta_1,\beta_2,\ldots,\beta_m) = Y_i - f(\underset{\sim}{x}_i;\ \beta_1,\ldots,\beta_m)$; for fixed $\beta_1,\ldots,\beta_m$ the d_i, $i = 1(1)n$, represent the deviations of the regression model with parameters $\beta_1,\ldots,\beta_m$ from the observed values $Y_1,\ldots,Y_n$.

The "ordinary" least squares principle would call for an estimator of $\underset{\sim}{\beta} = (\beta_1,\ldots,\beta_m)$ as that (unique) m-tuple of functions of $(\underset{\sim}{x}_1,\ldots,\underset{\sim}{x}_n)$, $\hat{\underset{\sim}{\beta}}^* = (\hat{\beta}_1^*,\ldots,\hat{\beta}_m^*)$, such that

$$\sum_1^n d_i^2(\hat{\underset{\sim}{\beta}}^*) = \min_{\underset{\sim}{\beta}} \sum_1^n d_i^2(\underset{\sim}{\beta})\ .$$

For some applications one might prefer to define the "weighted" least square estimator as that m-tuple of functions of $(x_1,\ldots,x_n)$, $\hat{\underset{\sim}{\beta}} = (\hat{\beta}_1,\ldots,\hat{\beta}_m)$, such that

$$\sum_1^n m_i d_i^2(\hat{\beta}) = \min_{\underset{\sim}{\beta}} \sum_1^n m_i d_i^2(\underset{\sim}{\beta})$$

where $m_i = m(\underset{\sim}{x}_i)$, m being a known function which weights the squared deviations depending on $\underset{\sim}{x}$ only. This last approach might be a reasonable one to take in estimating $\underset{\sim}{\beta}$ where ϵ is distributed with (known) variance $\sigma^2(\underset{\sim}{x})$ depending on $\underset{\sim}{x}$; in this case one might choose

$$m(x) = \frac{1}{\sigma^2(\underset{\sim}{x})} .$$

One can obtain an expression for $\hat{\beta}$ for a given sample $Y_1,\ldots,Y_n$ for fixed $\underset{\sim}{x}$ in the usual way:

Define
$$D(\underset{\sim}{\beta}) \equiv \sum_1^n m_i d_i^2(\underset{\sim}{\beta})$$

$$= \sum_1^n m_i [Y_i - f(\underset{\sim}{x}_i;\ \beta_1,\ldots,\beta_m)]^2 .$$

$$\frac{\partial D}{\partial \beta_i} = 2 \sum_1^n m_i [Y_i - f(x_i;\ \beta_1,\ldots,\beta_m)] \frac{\partial}{\partial \beta_i} f(\underset{\sim}{x}_i;\ \beta_1,\ldots,\beta_m)$$

$$i = 1(1)m .$$

$f(\underset{\sim}{x};\underset{\sim}{\beta})$ linear in $\underset{\sim}{\beta}$

$\Rightarrow \exists \{f_i(\underset{\sim}{x};\beta) \mid i = 1(1)m\}$

$\ni.\ f(\underset{\sim}{x};\underset{\sim}{\beta}) = \sum_1^m \beta_i f_i(\underset{\sim}{x})$ all $\underset{\sim}{x}$.

So
$$\frac{\partial D}{\partial \beta_i} = -2 \sum_{i=1}^n m_i \left[Y_i - \sum_{k=1}^m \beta_k f_k(\underset{\sim}{x}_i) \right] f_i(\underset{\sim}{x}_i)$$

$$i = 1(1)m$$

Let us write this system of equations as

$$\underset{\sim}{A\beta} = \underset{\sim}{C} \qquad (1)$$

with

$$A = \begin{bmatrix} \sum_1^n m_i f_1^2(\underset{\sim}{x}_i) & \sum_1^n m_i f_1(\underset{\sim}{x}_i) f_2(\underset{\sim}{x}_i) & \cdots & \sum_1^n m_i f_1(\underset{\sim}{x}_i) f_m(\underset{\sim}{x}_i) \\ & \sum_1^n m_i f_2^2(\underset{\sim}{x}_i) & \cdots & \sum_1^n m_i f_2(\underset{\sim}{x}_i) f_m(\underset{\sim}{x}_i) \\ & & & \vdots \\ \text{sym} & & & \sum_1^n m_i f_m^2(\underset{\sim}{x}_i) \end{bmatrix}$$

and

$$\underset{\sim}{C} = \begin{bmatrix} \sum_1^n m_i Y_i f_1(\underset{\sim}{x}_i) \\ \sum_1^n m_i Y_i f_2(\underset{\sim}{x}_i) \\ \vdots \\ \sum_1^n m_i Y_i f_m(\underset{\sim}{x}_i) \end{bmatrix}$$

and assume $|A| \neq 0$. $\hat{\underset{\sim}{\beta}}$ is the solution of (1).

The ordinary least-squares principle would result in the same equations (1) but with m_i, $i = 1(1)n$, replaced everywhere by 1. SAS provides for the fitting of equations by ordinary least-squares but not directly of weighted least-squares. However, of course one can accomplish the same results not by fitting the dependent variable to the independent ones with a weighting function but by fitting a function of the dependent variable to a function of the independent ones without. For, let

$$Y_i^o = \sqrt{m_i}\, Y_i \text{ with } m_i = m(\underset{\sim}{x}_i)$$

and $f_i^o(\underset{\sim}{x}) = \sqrt{m(x_i)}\, f_i(\underset{\sim}{x})$, $i = 1(1)m$; and substitute Y_i^o for Y_i, $f_i^o(\underset{\sim}{x})$ for $f_i(\underset{\sim}{x})$ and 1 for m_i in (1). One obtains the system of equations (1). We can use this simple fact to make SAS fit weighted linear regressions.

Example: Suppose one wants to fit the model

$$E(Y) = \beta_0 + \beta_1 \exp(-x_1) + \beta_2 x_1 x_2$$

where Var $(Y) = x_2^2\sigma^2$, σ^2 being an unknown constant. In our earlier notation we let

$$f_1(x_1,x_2) = 1$$
$$f_2(x_1,x_2) = \exp(-x_1)$$
$$f_3(x_1,x_2) = x_1x_2 \ .$$

To compensate for the heteroscedasticity of error variance let $m(x_1,x_2) = \frac{1}{x_2^2}$ In order to fit by OLS we let

$$f_1^o(x_1,x_2) = 1/|x_2|$$
$$f_2^o(x_1,x_2) = \exp(-x_1)/|x_2|$$
$$f_3^o(x_1,x_2) = x_1x_2/|x_2|$$

and $Y^o = Y/|x_2|$.

Note that if we regress Y^o on $f_1^o(x_1,x_2)$, $f_2^o(x_1,x_2)$ and $f_3^o(x_1,x_2)$ according to the model

$$Y^o = \beta_0 f_1^o(x_1,x_2) + \beta_1 f_2^o(x_1,x_2) + \beta_2 f_3^o(x_1,x_2) \tag{2}$$

the LS estimator of $(\beta_0,\beta_1,\beta_2)$ is the same as that for the weighted model. Moreover, var $(Y^o) = \frac{1}{x_2^2}$ var $(Y) = \sigma^2$ so that the residual MS in the model (2) is an estimator of σ^2 in the original model.

One can read the sample and fit the model by this scheme in SAS using the following program segment (see the SAS manual for explanation of programming details):

```
    .
    .
    .
comment read sample values;
input Y 1-2 1 x1 3-6 2 x2 10-14 2;
comment accomplish the transformations assuming that there are no missing
    values;
    f1_knot = 1/abs(x2);
    f2_knot = exp(-x1) * f1_knot;
    f3_knot = x1 * x2 * f1_knot;
    Y_knot = Y * f1_knot;
comment Y_knot indeed;
cards;
    < data cards >
proc regr;
comment fit a model with no explicit intercept term since the fitted coefficient
    for f1_knot is that value which would be obtained for the intercept were
    the original model fitted by weighted least squares;
model Y_knot = f1_knot
               f2_knot
               f3_knot/noint;
    .
    .
    .
```

Of course there are numerous other programs which facilitate the fitting of weighted linear regressions but SAS does have a number of file-processing features, etc which may conspire with the existence of this simple technique to make SAS seem more advantageous in some situations.